Best-Arm Identification in Linear Bandits

Marta Soare, Alessandro Lazaric, Rémi Munos

Best-Arm Identification in Linear Bandits

Marta Soare **Alessandro Lazaric** **Rémi Munos**[*] [†]
INRIA Lille – Nord Europe, SequeL Team
{marta.soare,alessandro.lazaric,remi.munos}@inria.fr

Abstract

We study the best-arm identification problem in linear bandit, where the rewards of the arms depend linearly on an unknown parameter θ^* and the objective is to return the arm with the largest reward. We characterize the complexity of the problem and introduce sample allocation strategies that pull arms to identify the best arm with a fixed confidence, while minimizing the sample budget. In particular, we show the importance of exploiting the global linear structure to improve the estimate of the reward of near-optimal arms. We analyze the proposed strategies and compare their empirical performance. Finally, as a by-product of our analysis, we point out the connection to the G-optimality criterion used in optimal experimental design.

1 Introduction

The stochastic multi-armed bandit problem (MAB) [16] offers a simple formalization for the study of sequential design of experiments. In the standard model, a learner sequentially chooses an arm out of K and receives a reward drawn from a fixed, unknown distribution relative to the chosen arm. While most of the literature in bandit theory focused on the problem of maximization of cumulative rewards, where the learner needs to trade-off exploration and exploitation, recently the *pure exploration* setting [5] has gained a lot of attention. Here, the learner uses the available budget to identify as accurately as possible the best arm, without trying to maximize the sum of rewards. Although many results are by now available in a wide range of settings (e.g., best-arm identification with fixed budget [2, 11] and fixed confidence [7], subset selection [6, 12], and multi-bandit [9]), most of the work considered only the multi-armed setting, with K independent arms.

An interesting variant of the MAB setup is the stochastic *linear bandit* problem (LB), introduced in [3]. In the LB setting, the input space $\mathcal{X}$ is a subset of $\mathbb{R}^d$ and when pulling an arm x, the learner observes a reward whose expected value is a linear combination of x and an unknown parameter $\theta^* \in \mathbb{R}^d$. Due to the linear structure of the problem, pulling an arm gives information about the parameter θ^* and indirectly, about the value of other arms. Therefore, the estimation of K mean-rewards is replaced by the estimation of the d features of θ^*. While in the exploration-exploitation setting the LB has been widely studied both in theory and in practice (e.g., [1, 14]), in this paper we focus on the pure-exploration scenario.

The fundamental difference between the MAB and the LB best-arm identification strategies stems from the fact that in MAB an arm is no longer pulled as soon as its sub-optimality is evident (in high probability), while in the LB setting even a sub-optimal arm may offer valuable information about the parameter vector θ^* and thus improve the accuracy of the estimation in discriminating among near-optimal arms. For instance, consider the situation when $K-2$ out of K arms are already discarded. In order to identify the best arm, MAB algorithms would concentrate the sampling on the two remaining arms to increase the accuracy of the estimate of their mean-rewards until the discarding condition is met for one of them. On the contrary, a LB pure-exploration strategy would seek to pull the arm $x \in \mathcal{X}$ whose observed reward allows to refine the estimate θ^* along the dimensions which are more suited in discriminating between the two remaining arms. Recently, the best-arm identification in linear bandits has been studied in a fixed budget setting [10], in this paper we study the sample complexity required to identify the best-linear arm with a fixed confidence.

[*] This work was done when the author was a visiting researcher at Microsoft Research New-England.
[†] Current affiliation: Google DeepMind.

2 Preliminaries

The setting. We consider the standard linear bandit model. Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a finite set of arms, where $|\mathcal{X}| = K$ and the ℓ_2-norm of any arm $x \in \mathcal{X}$, denoted by $||x||$, is upper-bounded by L. Given an unknown parameter $\theta^* \in \mathbb{R}^d$, we assume that each time an arm $x \in \mathcal{X}$ is pulled, a random reward $r(x)$ is generated according to the linear model $r(x) = x^\top \theta^* + \varepsilon$, where ε is a zero-mean i.i.d. noise bounded in $[-\sigma; \sigma]$. Arms are evaluated according to their expected reward $x^\top \theta^*$ and we denote by $x^* = \arg\max_{x \in \mathcal{X}} x^\top \theta^*$ the best arm in $\mathcal{X}$. Also, we use $\Pi(\theta) = \arg\max_{x \in \mathcal{X}} x^\top \theta$ to refer to the best arm corresponding to an arbitrary parameter θ. Let $\Delta(x, x') = (x - x')^\top \theta^*$ be the value *gap* between two arms, then we denote by $\Delta(x) = \Delta(x^*, x)$ the gap of x w.r.t. the optimal arm and by $\Delta_{\min} = \min_{x \in \mathcal{X}} \Delta(x)$ the minimum gap, where $\Delta_{\min} > 0$. We also introduce the sets $\mathcal{Y} = \{y = x - x', \forall x, x' \in \mathcal{X}\}$ and $\mathcal{Y}^* = \{y = x^* - x, \forall x \in \mathcal{X}\}$ containing all the directions obtained as the difference of two arms (or an arm and the optimal arm) and we redefine accordingly the gap of a direction as $\Delta(y) = \Delta(x, x')$ whenever $y = x - x'$.

The problem. We study the best-arm identification problem. Let $\hat{x}(n)$ be the estimated best arm returned by a bandit algorithm after n steps. We evaluate the *quality* of $\hat{x}(n)$ by the simple regret $R_n = (x^* - \hat{x}(n))^\top \theta^*$. While different settings can be defined (see [8] for an overview), here we focus on the (ϵ, δ)-best-arm identification problem (the so-called PAC setting), where given ϵ and $\delta \in (0, 1)$, the objective is to design an allocation strategy and a stopping criterion so that when the algorithm stops, the returned arm $\hat{x}(n)$ is such that $\mathbb{P}\big(R_n \geq \epsilon\big) \leq \delta$, while minimizing the needed number of steps. More specifically, we will focus on the case of $\epsilon = 0$ and we will provide high-probability bounds on the sample complexity n.

The multi-armed bandit case. In MAB, the complexity of best-arm identification is characterized by the gaps between arm values, following the intuition that the more similar the arms, the more pulls are needed to distinguish between them. More formally, the complexity is given by the problem-dependent quantity $H_{\text{MAB}} = \sum_{i=1}^{K} \frac{1}{\Delta_i^2}$ i.e., the inverse of the pairwise gaps between the best arm and the suboptimal arms. In the fixed budget case, H_{MAB} determines the probability of returning the wrong arm [2], while in the fixed confidence case, it characterizes the sample complexity [7].

Technical tools. Unlike in the multi-arm bandit scenario where pulling one arm does not provide any information about other arms, in a linear model we can leverage the rewards observed over time to estimate the expected reward of all the arms in $\mathcal{X}$. Let $\mathbf{x}_n = (x_1, \dots, x_n) \in \mathcal{X}^n$ be a sequence of arms and $(r_1, \dots, r_n)$ the corresponding observed (random) rewards. An unbiased estimate of θ^* can be obtained by ordinary least-squares (OLS) as $\hat{\theta}_n = A_{\mathbf{x}_n}^{-1} b_{\mathbf{x}_n}$, where $A_{\mathbf{x}_n} = \sum_{t=1}^{n} x_t x_t^\top \in \mathbb{R}^{d \times d}$ and $b_{\mathbf{x}_n} = \sum_{t=1}^{n} x_t r_t \in \mathbb{R}^d$. For any fixed sequence $\mathbf{x}_n$, through Azuma's inequality, the prediction error of the OLS estimate is upper-bounded in high-probability as follows.

Proposition 1. *Let $c = 2\sigma\sqrt{2}$ and $c' = 6/\pi^2$. For every fixed sequence $\mathbf{x}_n$, we have*[1]

$$\mathbb{P}\left(\forall n \in \mathbb{N}, \forall x \in \mathcal{X}, \left|x^\top \theta^* - x^\top \hat{\theta}_n\right| \leq c||x||_{A_{\mathbf{x}_n}^{-1}} \sqrt{\log(c' n^2 K/\delta)}\right) \geq 1 - \delta. \tag{1}$$

While in the previous statement $\mathbf{x}_n$ is fixed, a bandit algorithm adapts the allocation in response to the rewards observed over time. In this case a different high-probability bound is needed.

Proposition 2 (Thm. 2 in [1])**.** *Let $\hat{\theta}_n^\eta$ be the solution to the regularized least-squares problem with regularizer η and let $\widetilde{A}_{\mathbf{x}}^\eta = \eta I_d + A_{\mathbf{x}}$. Then for all $x \in \mathcal{X}$ and every adaptive sequence $\mathbf{x}_n$ such that at any step t, x_t only depends on $(x_1, r_1, \dots, x_{t-1}, r_{t-1})$, w.p. $1 - \delta$, we have*

$$\left|x^\top \theta^* - x^\top \hat{\theta}_n^\eta\right| \leq ||x||_{(\widetilde{A}_{\mathbf{x}_n}^\eta)^{-1}} \left(\sigma \sqrt{d \log\left(\frac{1 + nL^2/\eta}{\delta}\right)} + \eta^{1/2} ||\theta^*||\right). \tag{2}$$

The crucial difference w.r.t. Eq. 1 is an additional factor $\sqrt{d}$, the price to pay for adapting $\mathbf{x}_n$ to the samples. In the sequel we will often resort to the notion of design (or "soft" allocation) $\lambda \in \mathcal{D}^k$, which prescribes the *proportions* of pulls to arm x and $\mathcal{D}^k$ denotes the simplex $\mathcal{X}$. The counterpart of the design matrix A for a design λ is the matrix $\Lambda_\lambda = \sum_{x \in \mathcal{X}} \lambda(x) x x^\top$. From an allocation $\mathbf{x}_n$ we can derive the corresponding design $\lambda_{\mathbf{x}_n}$ as $\lambda_{\mathbf{x}_n}(x) = T_n(x)/n$, where $T_n(x)$ is the number of times arm x is selected in $\mathbf{x}_n$, and the corresponding design matrix is $A_{\mathbf{x}_n} = n\Lambda_{\lambda_{\mathbf{x}_n}}$.

[1] Whenever Prop.1 is used for all directions $y \in \mathcal{Y}$, then the logarithmic term becomes $\log(c' n^2 K^2/\delta)$ because of an additional union bound. For the sake of simplicity, in the sequel we always use $\log_n(K^2/\delta)$.

3 The Complexity of the Linear Best-Arm Identification Problem

As reviewed in Sect. 2, in the MAB case the complexity of the best-arm identification task is characterized by the reward gaps between the optimal and suboptimal arms. In this section, we propose an extension of the notion of complexity to the case of linear best-arm identification. In particular, we characterize the complexity by the performance of an *oracle* with access to the parameter θ^*.

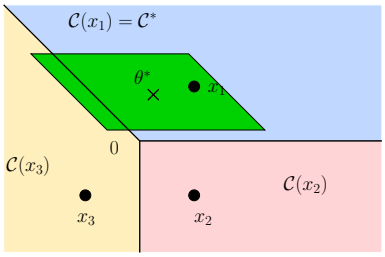

Figure 1: The cones corresponding to three arms (dots) in $\mathbb{R}^2$. Since $\theta^* \in \mathcal{C}(x_1)$, then $x^* = x_1$. The confidence set $\mathcal{S}^*(\mathbf{x}_n)$ (in green) is aligned with directions $x_1 - x_2$ and $x_1 - x_3$. Given the uncertainty in $\mathcal{S}^*(\mathbf{x}_n)$, both x_1 and x_3 may be optimal.

Stopping condition. Let $\mathcal{C}(x) = \{\theta \in \mathbb{R}^d, x \in \Pi(\theta)\}$ be the set of parameters θ which admit x as an optimal arm. As illustrated in Fig. 1, $\mathcal{C}(x)$ is the cone defined by the intersection of half-spaces such that $\mathcal{C}(x) = \cap_{x' \in \mathcal{X}} \{\theta \in \mathbb{R}^d, (x - x')^\top \theta \geq 0\}$ and all the cones together form a partition of the Euclidean space $\mathbb{R}^d$. We assume that the oracle knows the cone $\mathcal{C}(x^*)$ containing all the parameters for which x^* is optimal. Furthermore, we assume that for any allocation $\mathbf{x}_n$, it is possible to construct a confidence set $\mathcal{S}^*(\mathbf{x}_n) \subseteq \mathbb{R}^d$ such that $\theta^* \in \mathcal{S}^*(\mathbf{x}_n)$ and the (random) OLS estimate $\hat{\theta}_n$ belongs to $\mathcal{S}^*(\mathbf{x}_n)$ with high probability, i.e., $\mathbb{P}\big(\hat{\theta}_n \in \mathcal{S}^*(\mathbf{x}_n)\big) \geq 1 - \delta$. As a result, the oracle stopping criterion simply checks whether the confidence set $\mathcal{S}^*(\mathbf{x}_n)$ is contained in $\mathcal{C}(x^*)$ or not. In fact, whenever for an allocation $\mathbf{x}_n$ the set $\mathcal{S}^*(\mathbf{x}_n)$ overlaps the cones of different arms $x \in \mathcal{X}$, there is ambiguity in the identity of the arm $\Pi(\hat{\theta}_n)$. On the other hand when all possible values of $\hat{\theta}_n$ are included with high probability in the "right" cone $C(x^*)$, then the optimal arm is returned.

Lemma 1. *Let $\mathbf{x}_n$ be an allocation such that $\mathcal{S}^*(\mathbf{x}_n) \subseteq \mathcal{C}(x^*)$. Then $\mathbb{P}\big(\Pi(\hat{\theta}_n) \neq x^*\big) \leq \delta$.*

Arm selection strategy. From the previous lemma[2] it follows that the objective of an arm selection strategy is to define an allocation $\mathbf{x}_n$ which leads to $\mathcal{S}^*(\mathbf{x}_n) \subseteq \mathcal{C}(x^*)$ as quickly as possible.[3] Since this condition only depends on deterministic objects ($\mathcal{S}^*(\mathbf{x}_n)$ and $\mathcal{C}(x^*)$), it can be computed independently from the actual reward realizations. From a geometrical point of view, this corresponds to choosing arms so that the confidence set $\mathcal{S}^*(\mathbf{x}_n)$ shrinks into the optimal cone $\mathcal{C}(x^*)$ within the smallest number of pulls. To characterize this strategy we need to make explicit the form of $\mathcal{S}^*(\mathbf{x}_n)$. Intuitively speaking, the more $\mathcal{S}^*(\mathbf{x}_n)$ is "aligned" with the boundaries of the cone, the easier it is to shrink it into the cone. More formally, the condition $\mathcal{S}^*(\mathbf{x}_n) \subseteq \mathcal{C}(x^*)$ is equivalent to

$$\forall x \in \mathcal{X}, \forall \theta \in \mathcal{S}^*(\mathbf{x}_n), (x^* - x)^\top \theta \geq 0 \quad \Leftrightarrow \quad \forall y \in \mathcal{Y}^*, \forall \theta \in \mathcal{S}^*(\mathbf{x}_n), y^\top(\theta^* - \theta) \leq \Delta(y).$$

Then we can simply use Prop. 1 to directly control the term $y^\top(\theta^* - \theta)$ and define

$$\mathcal{S}^*(\mathbf{x}_n) = \left\{\theta \in \mathbb{R}^d, \forall y \in \mathcal{Y}^*, y^\top(\theta^* - \theta) \leq c||y||_{A_{\mathbf{x}_n}^{-1}} \sqrt{\log_n(K^2/\delta)}\right\}. \tag{3}$$

Thus the stopping condition $\mathcal{S}^*(\mathbf{x}_n) \subseteq \mathcal{C}(x^*)$ is equivalent to the condition that, for any $y \in \mathcal{Y}^*$,

$$c||y||_{A_{\mathbf{x}_n}^{-1}} \sqrt{\log_n(K^2/\delta)} \leq \Delta(y). \tag{4}$$

From this condition, the oracle allocation strategy simply follows as

$$\mathbf{x}_n^* = \arg\min_{\mathbf{x}_n} \max_{y \in \mathcal{Y}^*} \frac{c||y||_{A_{\mathbf{x}_n}^{-1}} \sqrt{\log_n(K^2/\delta)}}{\Delta(y)} = \arg\min_{\mathbf{x}_n} \max_{y \in \mathcal{Y}^*} \frac{||y||_{A_{\mathbf{x}_n}^{-1}}}{\Delta(y)}. \tag{5}$$

Notice that this strategy does not return an uniformly accurate estimate of θ^* but it rather pulls arms that allow to reduce the uncertainty of the estimation of θ^* over the directions of interest (i.e., $\mathcal{Y}^*$) below their corresponding gaps. This implies that the objective of Eq. 5 is to exploit the global linear assumption by pulling any arm in $\mathcal{X}$ that could give information about θ^* over the directions in $\mathcal{Y}^*$, so that directions with small gaps are better estimated than those with bigger gaps.

[2] For all the proofs in this paper, we refer the reader to the long version of the paper [18].

[3] Notice that by definition of the confidence set and since $\theta_n \to \theta^*$ as $n \to \infty$, any strategy repeatedly pulling all the arms would eventually meet the stopping condition.

Sample complexity. We are now ready to define the sample complexity of the oracle, which corresponds to the minimum number of steps needed by the allocation in Eq. 5 to achieve the stopping condition in Eq. 4. From a technical point of view, it is more convenient to express the complexity of the problem in terms of the optimal design (soft allocation) instead of the discrete allocation $\mathbf{x}_n$. Let $\rho^*(\lambda) = \max_{y \in \mathcal{Y}^*} ||y||^2_{\Lambda_\lambda^{-1}} / \Delta^2(y)$ be the square of the objective function in Eq. 5 for any design $\lambda \in \mathcal{D}^k$. We define the complexity of a linear best-arm identification problem as the performance achieved by the optimal design $\lambda^* = \arg\min_\lambda \rho^*(\lambda)$, i.e.

$$H_{\mathrm{LB}} = \min_{\lambda \in \mathcal{D}^k} \max_{y \in \mathcal{Y}^*} \frac{||y||^2_{\Lambda_\lambda^{-1}}}{\Delta^2(y)} = \rho^*(\lambda^*). \tag{6}$$

This definition of complexity is less explicit than in the case of H_{MAB} but it contains similar elements, notably the inverse of the gaps squared. Nonetheless, instead of summing the inverses over all the arms, H_{LB} implicitly takes into consideration the correlation between the arms in the term $||y||^2_{\Lambda_\lambda^{-1}}$, which represents the uncertainty in the estimation of the gap between x^* and x (when $y = x^* - x$). As a result, from Eq. 4 the sample complexity becomes

$$N^* = c^2 H_{\mathrm{LB}} \log_n(K^2/\delta), \tag{7}$$

where we use the fact that, if implemented over n steps, λ^* induces a design matrix $A_{\lambda^*} = n\Lambda_{\lambda^*}$ and $\max_y ||y||^2_{A_{\lambda^*}^{-1}} / \Delta^2(y) = \rho^*(\lambda^*)/n$. Finally, we bound the range of the complexity.

Lemma 2. *Given an arm set $\mathcal{X} \subseteq \mathbb{R}^d$ and a parameter θ^*, the complexity H_{LB} (Eq. 6) is such that*

$$\max_{y \in \mathcal{Y}^*} ||y||^2 / (L\Delta^2_{\min}) \leq H_{LB} \leq 4d/\Delta^2_{\min}. \tag{8}$$

Furthermore, if $\mathcal{X}$ is the canonical basis, the problem reduces to a MAB and $H_{MAB} \leq H_{LB} \leq 2H_{MAB}$.

The previous bounds show that $\Delta_{\min}$ plays a significant role in defining the complexity of the problem, while the specific shape of $\mathcal{X}$ impacts the numerator in different ways. In the worst case the full dimensionality d appears (upper-bound), and more arm-set specific quantities, such as the norm of the arms L and of the directions $\mathcal{Y}^*$, appear in the lower-bound.

4 Static Allocation Strategies

The oracle stopping condition (Eq. 4) and allocation strategy (Eq. 5) cannot be implemented in practice since θ^*, the gaps $\Delta(y)$, and the directions $\mathcal{Y}^*$ are unknown. In this section we investigate how to define algorithms that only rely on the information available from $\mathcal{X}$ and the samples collected over time. We introduce an empirical stopping criterion and two static allocations.

Input: decision space $\mathcal{X} \in \mathbb{R}^d$, confidence $\delta > 0$
Set: $t = 0$; $Y = \{y = (x - x'); x \neq x' \in \mathcal{X}\}$;
while Eq. 11 is not true **do**
  if G-allocation **then**
    $x_t = \arg\min_{x \in X} \max_{x' \in X} x'^\top (A + xx^\top)^{-1} x'$
  else if $\mathcal{XY}$-allocation **then**
    $x_t = \arg\min_{x \in X} \max_{y \in Y} y^\top (A + xx^\top)^{-1} y$
  end if
  Update $\hat{\theta}_t = A_t^{-1} b_t$, $t = t + 1$
end while
Return arm $\Pi(\hat{\theta}_t)$

Figure 2: Static allocation algorithms

Empirical stopping criterion. The stopping condition $\mathcal{S}^*(\mathbf{x}_n) \subseteq \mathcal{C}(x^*)$ cannot be tested since $\mathcal{S}^*(\mathbf{x}_n)$ is centered in the unknown parameter θ^* and $\mathcal{C}(x^*)$ depends on the unknown optimal arm x^*. Nonetheless, we notice that given $\mathcal{X}$, for each $x \in \mathcal{X}$ the cones $\mathcal{C}(x)$ can be constructed beforehand. Let $\widehat{\mathcal{S}}(\mathbf{x}_n)$ be a high-probability confidence set such that for any $\mathbf{x}_n$, $\hat{\theta}_n \in \widehat{\mathcal{S}}(\mathbf{x}_n)$ and $\mathbb{P}(\theta^* \in \widehat{\mathcal{S}}(\mathbf{x}_n)) \geq 1 - \delta$. Unlike $\mathcal{S}^*$, $\widehat{\mathcal{S}}$ can be directly computed from samples and we can stop whenever there exists an x such that $\widehat{\mathcal{S}}(\mathbf{x}_n) \subseteq \mathcal{C}(x)$.

Lemma 3. *Let $\mathbf{x}_n = (x_1, \dots, x_n)$ be an arbitrary allocation sequence. If after n steps there exists an arm $x \in \mathcal{X}$ such that $\widehat{\mathcal{S}}(\mathbf{x}_n) \subseteq \mathcal{C}(x)$ then $\mathbb{P}\big(\Pi(\hat{\theta}_n) \neq x^*\big) \leq \delta$.*

Arm selection strategy. Similarly to the oracle algorithm, we should design an allocation strategy that guarantees that the (random) confidence set $\widehat{\mathcal{S}}(\mathbf{x}_n)$ shrinks in one of the cones $\mathcal{C}(x)$ within the fewest number of steps. Let $\widehat{\Delta}_n(x, x') = (x - x')^\top \hat{\theta}_n$ be the empirical gap between arms x, x'. Then the stopping condition $\widehat{\mathcal{S}}(\mathbf{x}_n) \subseteq \mathcal{C}(x)$ can be written as

$$\begin{aligned} &\exists x \in \mathcal{X}, \forall x' \in \mathcal{X}, \forall \theta \in \widehat{\mathcal{S}}(\mathbf{x}_n), (x - x')^\top \theta \geq 0 \\ &\qquad \Leftrightarrow \quad \exists x \in \mathcal{X}, \forall x' \in \mathcal{X}, \forall \theta \in \widehat{\mathcal{S}}(\mathbf{x}_n), (x - x')^\top (\hat{\theta}_n - \theta) \leq \widehat{\Delta}_n(x, x'). \end{aligned} \tag{9}$$

This suggests that the empirical confidence set can be defined as

$$\widehat{\mathcal{S}}(\mathbf{x}_n) = \left\{\theta \in \mathbb{R}^d, \forall y \in \mathcal{Y}, y^\top(\hat{\theta}_n - \theta) \leq c||y||_{A_{\mathbf{x}_n}^{-1}} \sqrt{\log_n(K^2/\delta)}\right\}. \tag{10}$$

Unlike $\mathcal{S}^*(\mathbf{x}_n)$, $\widehat{\mathcal{S}}(\mathbf{x}_n)$ is centered in $\hat{\theta}_n$ and it considers all directions $y \in \mathcal{Y}$. As a result, the stopping condition in Eq. 9 could be reformulated as

$$\exists x \in \mathcal{X}, \forall x' \in \mathcal{X}, c||x - x'||_{A_{\mathbf{x}_n}^{-1}} \sqrt{\log_n(K^2/\delta)} \leq \widehat{\Delta}_n(x, x'). \tag{11}$$

Although similar to Eq. 4, unfortunately this condition cannot be directly used to derive an allocation strategy. In fact, it is considerably more difficult to define a suitable allocation strategy to fit a random confidence set $\widehat{\mathcal{S}}$ into a cone $\mathcal{C}(x)$ for an x which is not known in advance. In the following we propose two allocations that try to achieve the condition in Eq. 11 as fast as possible by implementing a static arm selection strategy, while we present a more sophisticated adaptive strategy in Sect. 5. The general structure of the static allocations in summarized in Fig. 2.

G-Allocation Strategy. The definition of the G-allocation strategy directly follows from the observation that for any pair $(x, x') \in \mathcal{X}^2$ we have that $||x - x'||_{A_{\mathbf{x}_n}^{-1}} \leq 2\max_{x'' \in \mathcal{X}} ||x''||_{A_{\mathbf{x}_n}^{-1}}$. This suggests that an allocation minimizing $\max_{x \in \mathcal{X}} ||x||_{A_{\mathbf{x}_n}^{-1}}$ reduces an upper bound on the quantity tested in the stopping condition in Eq. 11. Thus, for any fixed n, we define the G-allocation as

$$\mathbf{x}_n^G = \arg\min_{\mathbf{x}_n} \max_{x \in \mathcal{X}} ||x||_{A_{\mathbf{x}_n}^{-1}}. \tag{12}$$

We notice that this formulation coincides with the standard G-optimal design (hence the name of the allocation) defined in experimental design theory [15, Sect. 9.2] to minimize the maximal mean-squared prediction error in linear regression. The G-allocation can be interpreted as the design that allows to estimate θ^* *uniformly well* over all the arms in $\mathcal{X}$. Notice that the G-allocation in Eq. 12 is well defined only for a fixed number of steps n and it cannot be directly implemented in our case, since n is unknown in advance. Therefore we have to resort to a more "incremental" implementation. In the experimental design literature a wide number of approximate solutions have been proposed to solve the NP-hard discrete optimization problem in Eq. 12 (see [4, 17] for some recent results and [18] for a more thorough discussion). For any approximate G-allocation strategy with performance no worse than a factor $(1 + \beta)$ of the optimal strategy $\mathbf{x}_n^G$, the sample complexity N^G is bounded as follows.

Theorem 1. *If the G-allocation strategy is implemented with a β-approximate method and the stopping condition in Eq. 11 is used, then*

$$\mathbb{P}\left[N^G \leq \frac{16c^2 d(1+\beta)\log_n(K^2/\delta)}{\Delta_{\min}^2} \wedge \Pi(\hat{\theta}_{N^G}) = x^*\right] \geq 1 - \delta. \tag{13}$$

Notice that this result matches (up to constants) the worst-case value of N^* given the upper bound on H_{LB}. This means that, although completely static, the G-allocation is already worst-case optimal.

***$\mathcal{XY}$*-Allocation Strategy.** Despite being worst-case optimal, G-allocation is minimizing a rather loose upper bound on the quantity used to test the stopping criterion. Thus, we define an alternative static allocation that targets the stopping condition in Eq. 11 more directly by reducing its left-hand-side for any possible direction in $\mathcal{Y}$. For any fixed n, we define the $\mathcal{XY}$-allocation as

$$\mathbf{x}_n^{\mathcal{XY}} = \arg\min_{\mathbf{x}_n} \max_{y \in \mathcal{Y}} ||y||_{A_{\mathbf{x}_n}^{-1}}. \tag{14}$$

$\mathcal{XY}$-allocation is based on the observation that the stopping condition in Eq. 11 requires only the empirical gaps $\widehat{\Delta}(x, x')$ to be well estimated, hence arms are pulled with the objective of increasing the accuracy of directions in $\mathcal{Y}$ instead of arms $\mathcal{X}$. This problem can be seen as a transductive variant of the G-optimal design [19], where the target vectors $\mathcal{Y}$ are different from the vectors $\mathcal{X}$ used in the design. The sample complexity of the $\mathcal{XY}$-allocation is as follows.

Theorem 2. *If the $\mathcal{XY}$-allocation strategy is implemented with a β-approximate method and the stopping condition in Eq. 11 is used, then*

$$\mathbb{P}\left[N^{\mathcal{XY}} \leq \frac{32c^2 d(1+\beta)\log_n(K^2/\delta)}{\Delta_{\min}^2} \wedge \Pi(\hat{\theta}_{N^{\mathcal{XY}}}) = x^*\right] \geq 1 - \delta. \tag{15}$$

Although the previous bound suggests that $\mathcal{XY}$ achieves a performance comparable to the G-allocation, in fact $\mathcal{XY}$ may be arbitrarily better than G-allocation (for an example, see [18]).

5 $\mathcal{XY}$-Adaptive Allocation Strategy

Fully adaptive allocation strategies. Although both G- and $\mathcal{XY}$-allocation are sound since they minimize upper-bounds on the quantities used by the stopping condition (Eq. 11), they may be very sub-optimal w.r.t. the ideal performance of the oracle introduced in Sec. 3. Typically, an improvement can be obtained by moving to strategies adapting on the rewards observed over time. Nonetheless, as reported in Prop. 2, whenever $\mathbf{x}_n$ is not a fixed sequence, the bound in Eq. 2 should be used. As a result, a factor $\sqrt{d}$ would appear in the definition of the confidence sets and in the stopping condition. This directly implies that the sample complexity of a fully adaptive strategy would scale linearly with the dimensionality d of the problem, thus removing any advantage w.r.t. static allocations. In fact, the sample complexity of G- and $\mathcal{XY}$-allocation already scales linearly with d and from Lem. 2 we cannot expect to improve the dependency on $\Delta_{\min}$. Thus, on the one hand, we need to use the tighter bounds in Eq. 1 and, on the other hand, we require to be adaptive w.r.t. samples. In the sequel we propose a phased algorithm which successfully meets both requirements using a static allocation within each phase but choosing the type of allocation depending on the samples observed in previous phases.

Input: decision space $\mathcal{X} \in \mathbb{R}^d$; parameter α; confidence δ
Set $j=1$; $\widehat{\mathcal{X}}_j = \mathcal{X}$; $\widehat{\mathcal{Y}}_1 = \mathcal{Y}$; $\rho_0 = 1$; $n_0 = d(d+1)+1$
while $|\widehat{\mathcal{X}}_j| > 1$ **do**
  $\rho^j = \rho^{j-1}$
  $t = 1$; $A_0 = I$
  while $\rho^j / t \geq \alpha \rho^{j-1}(\mathbf{x}_{n_{j-1}}^{j-1})/n_{j-1}$ **do**
    Select arm $x_t = \arg\min_{x \in X} \max_{y \in Y} y^\top (A + xx^\top)^{-1} y$
    Update $A_t = A_{t-1} + x_t x_t^\top$, $t = t+1$
    $\rho^j = \max_{y \in \widehat{\mathcal{Y}}_j} y^\top A_t^{-1} y$
  end while
  Compute $b = \sum_{s=1}^t x_s r_s$; $\hat{\theta}_j = A_t^{-1} b$
  $\widehat{\mathcal{X}}_{j+1} = \mathcal{X}$
  for $x \in \mathcal{X}$ **do**
    if $\exists x' : ||x - x'||_{A_t^{-1}} \sqrt{\log_n(K^2/\delta)} \leq \widehat{\Delta}_j(x', x)$ **then**
      $\widehat{\mathcal{X}}_{j+1} = \widehat{\mathcal{X}}_{j+1} - \{x\}$
    end if
  end for
  $\widehat{\mathcal{Y}}_{j+1} = \{y = (x - x'); x, x' \in \widehat{\mathcal{X}}_{j+1}\}$
end while
Return $\Pi(\hat{\theta}_j)$

Figure 3: $\mathcal{XY}$-Adaptive allocation algorithm

Algorithm. The ideal case would be to define an empirical version of the oracle allocation in Eq. 5 so as to adjust the accuracy of the prediction only on the directions of interest $\mathcal{Y}^*$ and according to their gaps $\Delta(y)$. As discussed in Sect. 4 this cannot be obtained by a direct adaptation of Eq. 11. In the following, we describe a safe alternative to adjust the allocation strategy to the gaps.

Lemma 4. *Let $\mathbf{x}_n$ be a fixed allocation sequence and $\hat{\theta}_n$ its corresponding estimate for θ^*. If an arm $x \in \mathcal{X}$ is such that*

$$\exists x' \in \mathcal{X} \text{ s.t. } c||x' - x||_{A_{\mathbf{x}_n}^{-1}} \sqrt{\log_n(K^2/\delta)} < \widehat{\Delta}_n(x', x), \tag{16}$$

then arm x is sub-optimal. Moreover, if Eq. 16 is true, we say that x' dominates x.

Lem. 4 allows to easily construct the set of potentially optimal arms, denoted $\widehat{\mathcal{X}}(\mathbf{x}_n)$, by removing from $\mathcal{X}$ all the dominated arms. As a result, we can replace the stopping condition in Eq. 11, by just testing whether the number of non-dominated arms $|\widehat{\mathcal{X}}(\mathbf{x}_n)|$ is equal to 1, which corresponds to the case where the confidence set is fully contained into a single cone. Using $\widehat{\mathcal{X}}(\mathbf{x}_n)$, we construct $\widehat{\mathcal{Y}}(\mathbf{x}_n) = \{y = x - x'; x, x' \in \widehat{\mathcal{X}}(\mathbf{x}_n)\}$, the set of directions along which the estimation of θ^* needs to be improved to further shrink $\widehat{\mathcal{S}}(\mathbf{x}_n)$ into a single cone and trigger the stopping condition. Note that if $\mathbf{x}_n$ was an adaptive strategy, then we could not use Lem. 4 to discard arms but we should rely on the bound in Prop. 2. To avoid this problem, an effective solution is to run the algorithm through phases. Let $j \in \mathbb{N}$ be the index of a phase and n_j its corresponding length. We denote by $\widehat{\mathcal{X}}_j$ the set of non-dominated arms constructed on the basis of the samples collected in the phase $j-1$. This set is used to identify the directions $\widehat{\mathcal{Y}}_j$ and to define a *static* allocation which focuses on reducing the uncertainty of θ^* along the directions in $\widehat{\mathcal{Y}}_j$. Formally, in phase j we implement the allocation

$$\mathbf{x}_{n_j}^j = \arg\min_{\mathbf{x}_{n_j}} \max_{y \in \widehat{\mathcal{Y}}_j} ||y||_{A_{\mathbf{x}_{n_j}}^{-1}}, \tag{17}$$

which coincides with a $\mathcal{XY}$-allocation (see Eq. 14) but restricted on $\widehat{\mathcal{Y}}_j$. Notice that $\mathbf{x}_{n_j}^j$ may still use any arm in $\mathcal{X}$ which could be useful in reducing the confidence set along any of the directions in

$\widehat{\mathcal{Y}}_j$. Once phase j is over, the OLS estimate $\hat{\theta}^j$ is computed using the rewards observed within phase j and then is used to test the stopping condition in Eq. 11. Whenever the stopping condition does not hold, a new set $\widehat{\mathcal{X}}_{j+1}$ is constructed using the discarding condition in Lem. 4 and a new phase is started. Notice that through this process, at each phase j, the allocation $\mathbf{x}_{n_j}^j$ is static conditioned on the previous allocations and the use of the bound from Prop. 1 is still correct.

A crucial aspect of this algorithm is the length of the phases n_j. On the one hand, short phases allow a high rate of adaptivity, since $\widehat{\mathcal{X}}_j$ is recomputed very often. On the other hand, if a phase is too short, it is very unlikely that the estimate $\hat{\theta}^j$ may be accurate enough to actually discard any arm. An effective way to define the length of a phase in a deterministic way is to relate it to the actual uncertainty of the allocation in estimating the value of all the active directions in $\widehat{\mathcal{Y}}_j$. In phase j, let $\rho^j(\lambda) = \max_{y \in \widehat{\mathcal{Y}}_j} ||y||^2_{\Lambda_\lambda^{-1}}$, then given a parameter $\alpha \in (0,1)$, we define

$$n_j = \min\left\{n \in \mathbb{N} : \rho^j(\lambda_{\mathbf{x}_n^j})/n \leq \alpha \rho^{j-1}(\lambda^{j-1})/n_{j-1}\right\}, \tag{18}$$

where $\mathbf{x}_n^j$ is the allocation defined in Eq. 17 and λ^{j-1} is the design corresponding to $\mathbf{x}_{n_{j-1}}^{j-1}$, the allocation performed at phase $j-1$. In words, n_j is the minimum number of steps needed by the $\mathcal{XY}$-adaptive allocation to achieve an uncertainty over all the directions of interest which is a fraction α of the performance obtained in the previous iteration. Notice that given $\widehat{\mathcal{Y}}_j$ and ρ^{j-1} this quantity can be computed before the actual beginning of phase j. The resulting algorithm using the $\mathcal{XY}$-Adaptive allocation strategy is summarized in Fig. 3.

Sample complexity. Although the $\mathcal{XY}$-Adaptive allocation strategy is designed to approach the oracle sample complexity N^*, in early phases it basically implements a $\mathcal{XY}$-allocation and no significant improvement can be expected until some directions are discarded from $\widehat{\mathcal{Y}}$. At that point, $\mathcal{XY}$-adaptive starts focusing on directions which only contain near-optimal arms and it starts approaching the behavior of the oracle. As a result, in studying the sample complexity of $\mathcal{XY}$-Adaptive we have to take into consideration the unavoidable price of discarding "suboptimal" directions. This cost is directly related to the geometry of the arm space that influences the number of samples needed before arms can be discarded from $\mathcal{X}$. To take into account this problem-dependent quantity, we introduce a slightly relaxed definition of complexity. More precisely, we define the number of steps needed to discard all the directions which do not contain x^*, i.e. $\mathcal{Y} - \mathcal{Y}^*$. From a geometrical point of view, this corresponds to the case when for any pair of suboptimal arms (x, x'), the confidence set $\mathcal{S}^*(\mathbf{x}_n)$ does not intersect the hyperplane separating the cones $\mathcal{C}(x)$ and $\mathcal{C}(x')$. Fig. 1 offers a simple illustration for such a situation: $\mathcal{S}^*$ no longer intercepts the border line between $\mathcal{C}(x_2)$ and $\mathcal{C}(x_3)$, which implies that direction $x_2 - x_3$ can be discarded. More formally, the hyperplane containing parameters θ for which x and x' are equivalent is simply $\mathcal{C}(x) \cap \mathcal{C}(x')$ and the quantity

$$M^* = \min\{n \in \mathbb{N}, \forall x \neq x^*, \forall x' \neq x^*, \mathcal{S}^*(\mathbf{x}_n^{\mathcal{XY}}) \cap (\mathcal{C}(x) \cap \mathcal{C}(x')) = \emptyset\} \tag{19}$$

corresponds to the minimum number of steps needed by the static $\mathcal{XY}$-allocation strategy to discard all the *suboptimal* directions. This term together with the oracle complexity N^* characterizes the sample complexity of the phases of the $\mathcal{XY}$-adaptive allocation. In fact, the length of the phases is such that either they correspond to the complexity of the oracle or they can never last more than the steps needed to discard all the sub-optimal directions. As a result, the overall sample complexity of the $\mathcal{XY}$-adaptive algorithm is bounded as in the following theorem.

Theorem 3. *If the $\mathcal{XY}$-Adaptive allocation strategy is implemented with a β-approximate method and the stopping condition in Eq. 11 is used, then*

$$\mathbb{P}\left[N \leq \frac{(1+\beta)\max\{M^*, \frac{16}{\alpha}N^*\}}{\log(1/\alpha)} \log\left(\frac{c\sqrt{\log_n(K^2/\delta)}}{\Delta_{\min}}\right) \wedge \Pi(\hat{\theta}_N) = x^*\right] \geq 1 - \delta. \tag{20}$$

We first remark that, unlike G and $\mathcal{XY}$, the sample complexity of $\mathcal{XY}$-Adaptive does not have any direct dependency on d and $\Delta_{\min}$ (except in the logarithmic term) but it rather scales with the oracle complexity N^* and the cost of discarding suboptimal directions M^*. Although this additional cost is probably unavoidable, one may have expected that $\mathcal{XY}$-Adaptive may need to discard all the suboptimal directions before performing as well as the oracle, thus having a sample complexity of $O(M^* + N^*)$. Instead, we notice that N scales with the *maximum* of M^* and N^*, thus implying that $\mathcal{XY}$-Adaptive may actually catch up with the performance of the oracle (with only a multiplicative factor of $16/\alpha$) whenever discarding suboptimal directions is less expensive than actually identifying the best arm.

6 Numerical Simulations

We illustrate the performance of $\mathcal{XY}$-Adaptive and compare it to the $\mathcal{XY}$-Oracle strategy (Eq. 5), the static allocations $\mathcal{XY}$ and G, as well as with the fully-adaptive version of $\mathcal{XY}$ where $\widehat{\mathcal{X}}$ is updated at each round and the bound from Prop.2 is used. For a fixed confidence $\delta = 0.05$, we compare the sampling budget needed to identify the best arm with probability at least $1 - \delta$. We consider a set of arms $\mathcal{X} \in \mathbb{R}^d$, with $|\mathcal{X}| = d + 1$ including the canonical basis $(e_1, \dots, e_d)$ and an additional arm $x_{\text{d+1}} = [\cos(\omega) \;\; \sin(\omega) \;\; 0 \;\; \dots \;\; 0]^\top$. We choose $\theta^* = [2 \;\; 0 \;\; 0 \;\; \dots \;\; 0]^\top$, and fix $\omega = 0.01$, so that $\Delta_{\min} = (x_1 - x_{\text{d+1}})^\top \theta^*$ is much smaller than the other gaps. In this setting, an efficient sampling strategy should focus on reducing the uncertainty in the direction $\tilde{y} = (x_1 - x_{\text{d+1}})$ by pulling the arm $x_2 = e_2$ which is almost aligned with $\tilde{y}$. In fact, from the rewards obtained from x_2 it is easier to decrease the uncertainty about the second component of θ^*, that is precisely the dimension which allows to discriminate between x_1 and $x_{\text{d+1}}$. Also, we fix $\alpha = 1/10$, and the noise $\varepsilon \sim \mathcal{N}(0, 1)$. Each phase begins with an initialization matrix A_0, obtained by pulling once each canonical arm. In Fig. 4 we report the sampling budget of the algorithms, averaged over 100 runs, for $d = 2 \dots 10$.

The results. The numerical results show that $\mathcal{XY}$-Adaptive is effective in allocating the samples to shrink the uncertainty in the direction $\tilde{y}$. Indeed, $\mathcal{XY}$-adaptive identifies the most important direction after few phases and is able to perform an allocation which mimics that of the oracle. On the contrary, $\mathcal{XY}$ and G do not adjust to the empirical gaps and consider all directions as equally important. This behavior forces $\mathcal{XY}$ and G to allocate samples until the uncertainty is smaller than $\Delta_{\min}$ in all directions. Even though the Fully-adaptive algorithm also identifies the most informative direction rapidly, the $\sqrt{d}$ term in the bound delays the discarding of the arms and prevents the algorithm from gaining any advantage compared to $\mathcal{XY}$ and G. As shown in Fig. 4, the difference between the budget of $\mathcal{XY}$-Adaptive and the static strategies increases with the number of dimensions. In fact, while additional dimensions have little to no impact on $\mathcal{XY}$-Oracle and $\mathcal{XY}$-Adaptive (the only important direction remains $\tilde{y}$ independently from the number of unknown features of θ^*), for the static allocations more dimensions imply more directions to be considered and more features of θ^* to be estimated uniformly well until the uncertainty falls below $\Delta_{\min}$.

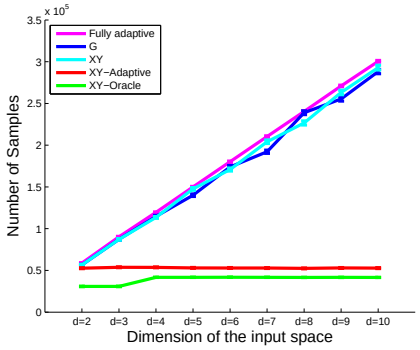

Figure 4: The sampling budget needed to identify the best arm, when the dimension grows from $\mathbb{R}^2$ to $\mathbb{R}^{10}$.

7 Conclusions

In this paper we studied the problem of best-arm identification with a fixed confidence, in the linear bandit setting. First we offered a preliminary characterization of the problem-dependent complexity of the best arm identification task and shown its connection with the complexity in the MAB setting. Then, we designed and analyzed efficient sampling strategies for this problem. The G-allocation strategy allowed us to point out a close connection with optimal experimental design techniques, and in particular to the G-optimality criterion. Through the second proposed strategy, $\mathcal{XY}$-allocation, we introduced a novel optimal design problem where the testing arms do not coincide with the arms chosen in the design. Lastly, we pointed out the limits that a fully-adaptive allocation strategy might have in the linear bandit setting and proposed a phased-algorithm, $\mathcal{XY}$-Adaptive, that learns from previous observations, without suffering from the dimensionality of the problem. Since this is one of the first works that analyze pure-exploration problems in the linear-bandit setting, it opens the way for an important number of similar problems already studied in the MAB setting. For instance, we can investigate strategies to identify the best-linear arm when having a limited budget or study the best-arm identification when the set of arms is very large (or infinite). Some interesting extensions also emerge from the optimal experimental design literature, such as the study of sampling strategies for meeting the G-optimality criterion when the noise is heterosckedastic, or the design of efficient strategies for satisfying other related optimality criteria, such as V-optimality.

Acknowledgments This work was supported by the French Ministry of Higher Education and Research, Nord-Pas de Calais Regional Council and FEDER through the “Contrat de Projets Etat Region 2007–2013", and European Community’s Seventh Framework Programme under grant agreement no 270327 (project CompLACS).

References

[1] Yasin Abbasi-Yadkori, Dávid Pál, and Csaba Szepesvári. Improved algorithms for linear stochastic bandits. In *Proceedings of the 25th Annual Conference on Neural Information Processing Systems (NIPS)*, 2011.

[2] Jean-Yves Audibert, Sébastien Bubeck, and Rémi Munos. Best arm identification in multi-armed bandits. In *Proceedings of the 23rd Conference on Learning Theory (COLT)*, 2010.

[3] Peter Auer. Using confidence bounds for exploitation-exploration trade-offs. *Journal of Machine Learning Research*, 3:397–422, 2002.

[4] Mustapha Bouhtou, Stephane Gaubert, and Guillaume Sagnol. Submodularity and randomized rounding techniques for optimal experimental design. *Electronic Notes in Discrete Mathematics*, 36:679–686, 2010.

[5] Sébastien Bubeck, Rémi Munos, and Gilles Stoltz. Pure exploration in multi-armed bandits problems. In *Proceedings of the 20th International Conference on Algorithmic Learning Theory (ALT)*, 2009.

[6] Sébastien Bubeck, Tengyao Wang, and Nitin Viswanathan. Multiple identifications in multi-armed bandits. In *Proceedings of the International Conference in Machine Learning (ICML)*, pages 258–265, 2013.

[7] Eyal Even-Dar, Shie Mannor, and Yishay Mansour. Action elimination and stopping conditions for the multi-armed bandit and reinforcement learning problems. *J. Mach. Learn. Res.*, 7:1079–1105, December 2006.

[8] Victor Gabillon, Mohammad Ghavamzadeh, and Alessandro Lazaric. Best arm identification: A unified approach to fixed budget and fixed confidence. In *Proceedings of the 26th Annual Conference on Neural Information Processing Systems (NIPS)*, 2012.

[9] Victor Gabillon, Mohammad Ghavamzadeh, Alessandro Lazaric, and Sébastien Bubeck. Multi-bandit best arm identification. In *Proceedings of the 25th Annual Conference on Neural Information Processing Systems (NIPS)*, pages 2222–2230, 2011.

[10] Matthew D. Hoffman, Bobak Shahriari, and Nando de Freitas. On correlation and budget constraints in model-based bandit optimization with application to automatic machine learning. In *Proceedings of the 17th International Conference on Artificial Intelligence and Statistics (AISTATS)*, pages 365–374, 2014.

[11] Kevin G. Jamieson, Matthew Malloy, Robert Nowak, and Sébastien Bubeck. lil' UCB : An optimal exploration algorithm for multi-armed bandits. In *Proceeding of the 27th Conference on Learning Theory (COLT)*, 2014.

[12] Emilie Kaufmann and Shivaram Kalyanakrishnan. Information complexity in bandit subset selection. In *Proceedings of the 26th Conference on Learning Theory (COLT)*, pages 228–251, 2013.

[13] Jack Kiefer and Jacob Wolfowitz. The equivalence of two extremum problems. *Canadian Journal of Mathematics*, 12:363–366, 1960.

[14] Lihong Li, Wei Chu, John Langford, and Robert E. Schapire. A contextual-bandit approach to personalized news article recommendation. In *Proceedings of the 19th International Conference on World Wide Web (WWW)*, pages 661–670, 2010.

[15] Friedrich Pukelsheim. *Optimal Design of Experiments*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, 2006.

[16] Herbert Robbins. Some aspects of the sequential design of experiments. *Bulletin of the American Mathematical Society*, pages 527–535, 1952.

[17] Guillaume Sagnol. Approximation of a maximum-submodular-coverage problem involving spectral functions, with application to experimental designs. *Discrete Appl. Math.*, 161(1-2):258–276, January 2013.

[18] Marta Soare, Alessandro Lazaric, and Rémi Munos. Best-Arm Identification in Linear Bandits. Technical report, http://arxiv.org/abs/1409.6110.

[19] Kai Yu, Jinbo Bi, and Volker Tresp. Active learning via transductive experimental design. In *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pages 1081–1088, 2006.

A Comparison between G-allocation and $\mathcal{XY}$-allocation

We define two examples illustrating the difference between the G and the $\mathcal{XY}$ allocation strategies. Let us consider a problem with $\mathcal{X} \subset \mathbb{R}^2$ and arms $x_1 = [1 \;\; \epsilon/2]^\top$ and $x_2 = [1 \;\; -\epsilon/2]^\top$, where $\epsilon \in (0, 1)$. In this case, both static allocations pull the two arms the same number of times, thus inducing an optimal design $\lambda(x_1) = \lambda(x_2) = 1/2$. We want to study the (asymptotic) performance of the allocation according to the different definition of error $\max_{x \in \mathcal{X}} x^\top \Lambda_\lambda^{-1} x$ and $\max_{y \in \mathcal{Y}} y^\top \Lambda_\lambda^{-1} y$ used by G and $\mathcal{XY}$-allocation respectively. We first notice that

$$\Lambda_\lambda = \frac{1}{2} \begin{bmatrix} 1 & \epsilon/2 \\ \epsilon/2 & \epsilon^2/4 \end{bmatrix} + \frac{1}{2} \begin{bmatrix} 1 & -\epsilon/2 \\ -\epsilon/2 & \epsilon^2/4 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & \epsilon^2/4 \end{bmatrix}.$$

As a result, for both x_1 and x_2 we have

$$[1 \;\; \epsilon/2] \Lambda_\lambda^{-1} \begin{bmatrix} 1 \\ \epsilon/2 \end{bmatrix} = [1 \;\; \epsilon/2] \begin{bmatrix} 1 & 0 \\ 0 & 4/\epsilon^2 \end{bmatrix} \begin{bmatrix} 1 \\ \epsilon/2 \end{bmatrix} = 2.$$

On the other hand, if we consider the direction $y = x_1 - x_2 = [0 \;\; \epsilon]$, we have

$$[0 \;\; \epsilon] \Lambda_\lambda^{-1} \begin{bmatrix} 0 \\ \epsilon \end{bmatrix} = [0 \;\; \epsilon] \begin{bmatrix} 1 & 0 \\ 0 & 4/\epsilon^2 \end{bmatrix} \begin{bmatrix} 0 \\ \epsilon \end{bmatrix} = 4.$$

This example shows that indeed the performance achieved by $\mathcal{XY}$ may be similar to the performance of G-optimal. Let us now consider a different setting where the two arms $x_1 = [1 \;\; 0]$ and $x_2 = [1 - \epsilon \;\; 0]$ are aligned on the same axis. In this case, the problem reduces to a 1-dimensional problem and both strategies would concentrate their allocation on $x_1 = [1 \;\; 0]$ since it is the arm with larger norm and it may provide a better estimate of θ^*. As a result, while the G-allocation has a performance of 1, the $\mathcal{XY}$-allocation over the direction $[\epsilon \;\; 0]$ has a performance ϵ^2, which can be arbitrarily smaller than 1.

B Proofs

B.1 Lemmas

Proof of Lemma 1. The proof follows from the fact that if $\mathcal{S}^*(\mathbf{x}_n) \subseteq \mathcal{C}(x^*)$ and $\hat{\theta}_n \in \mathcal{S}^*(\mathbf{x}_n)$ with high probability, then $\hat{\theta}_n \in \mathcal{C}(x^*)$ which implies that $\Pi(\hat{\theta}_n) = x^*$ by definition of the cone $\mathcal{C}(x^*)$. □

Before proceeding to the proof of Lemma 2 we introduce the following technical tool.

Proposition 3 (Equivalence-Theorem in [13])**.** *Define $f(x; \xi) = x^\top M(\xi)^{-1} x$, where $M(\xi)$ is a $d \times d$ non-singular matrix and x is a column vector in $\mathbb{R}^d$. We consider two extremum problems.*

The first is to choose ξ so that

$$(1) \;\; \xi \;\; \textit{maximizes} \;\; \det M(\xi) \qquad (\textit{D-optimal design})$$

The second one is to choose ξ so that

$$(2) \;\; \xi \;\; \textit{minimizes} \;\; \max f(x; \xi) \qquad (\textit{G-optimal design})$$

We note that the integral with respect to ξ of $f(x; \xi)$ is d; hence, $\max f(x; \xi) \geq d$, and thus a sufficient condition for ξ to satisfy (2) is

$$(3) \;\; \max f(x; \xi) = d.$$

Statements (1), (2) and (3) are equivalent.

Proof of Lemma 2. **Upper-bound.** We have the following sequence of inequalities

$$\max_{y \in \mathcal{Y}_*} \frac{||y||^2_{\Lambda_\lambda^{-1}}}{\Delta^2(y)} \leq \frac{1}{\Delta^2_{\min}} \max_{y \in \mathcal{Y}_*} ||y||^2_{\Lambda_\lambda^{-1}} \leq \frac{4}{\Delta^2_{\min}} \max_{x \in \mathcal{X}} ||x||^2_{\Lambda_\lambda^{-1}},$$

where the second inequality comes from a triangle inequality on $||y||^2_{\Lambda_\lambda^{-1}}$. Thus we obtain

$$\rho^*(\lambda^*) = \min_{\lambda\in\mathcal{D}^k} \max_{y\in\mathcal{Y}_*} \frac{||y||^2_{\Lambda_\lambda^{-1}}}{\Delta^2(y)} \leq \frac{4}{\Delta^2_{\min}} \min_{\lambda\in\mathcal{D}^k} \max_{x\in\mathcal{X}} ||x||^2_{\Lambda_\lambda^{-1}} = \frac{4d}{\Delta^2_{\min}},$$

where the last equality follows from the Kiefer-Wolfowitz equivalence theorem presented in Prop. 3.

Lower-bound.

We focus on the numerator $y^\top \Lambda_\lambda^{-1} y$. Since Λ_λ is a positive definite matrix, we define its decomposition $\Lambda_\lambda = Q\Gamma Q^\top$, where Q is an orthogonal matrix and Γ is the diagonal matrix containing the eigenvalues. As a result the numerator can be written as

$$y^\top \Lambda_\lambda^{-1} y = y^\top Q\Gamma^{-1} Q^\top y = w^\top \Gamma^{-1} w,$$

where we renamed $Q^\top y = w$. If we denote by $\gamma_{\max}$ the largest eigenvalue of Λ_λ (i.e., the largest value in Γ), then

$$w^\top \Gamma^{-1} w \geq 1/\gamma_{\max} w^\top w = 1/\gamma_{\max} ||y||^2.$$

The largest eigenvalue $\gamma_{\max}$ is upper-bounded by the sum of the largest eigenvalues of the matrices $\lambda(x)xx^\top$ which is $\lambda(x)||x||_2$. As a result, we obtain the bound $\gamma_{\max} \leq \sum_x \lambda(x)||x||_2 \leq L$, since $||x||_2 \leq L$ and λ is in the simplex. Thus we have

$$\min_{\lambda\in\mathcal{D}^k} \max_{y\in\mathcal{Y}_*} \frac{||y||^2_{\Lambda_\lambda^{-1}}}{\Delta^2(y)} \geq \frac{1}{L} \max_{y\in\mathcal{Y}^*} \frac{||y||^2}{\Delta(y)^2} \geq \frac{\max_{y\in\mathcal{Y}^*} ||y||^2}{L\Delta^2_{\min}}.$$

Comparison with the K-armed bandit complexity.

Finally, we show how the sample complexity reduces to the known quantity in the MAB case. If the arms in $\mathcal{X}$ coincide with the canonical basis of $\mathbb{R}^d$, then for any allocation λ the design matrix Λ_λ becomes a diagonal matrix of the form $\text{diag}(\lambda(x_1), \ldots, \lambda(x_K))$. As a result, we obtain

$$H_{\text{LB}} = \min_{\lambda\in\mathcal{D}^k} \max_{y\in\mathcal{Y}^*} \frac{||y||^2_{\Lambda_\lambda^{-1}}}{\Delta^2(y)} = \min_{\lambda\in\mathcal{D}^k} \max_{x\in\mathcal{X}-\{x^*\}} \frac{1/\lambda(x) + 1/\lambda(x^*)}{\Delta^2(x)}.$$

If we use the allocation $\lambda(x) = 1/(\nu\Delta^2(x))$ and $\lambda(x^*) = 1/(\nu\Delta_{\min})$, with $\nu = 1/\Delta^2_{\min} + \sum_{x\neq x^*} 1/\Delta^2(x)$, we obtain

$$\begin{aligned} H_{\text{LB}} &\leq \max_{x\in\mathcal{X}-\{x^*\}} \frac{\nu\Delta^2(x) + \nu\Delta^2_{\min}}{\Delta^2(x)} = \max_{x\in\mathcal{X}-\{x^*\}} \nu + \nu\frac{\Delta^2_{\min}}{\Delta^2(x)} \\ &= 2\nu = 2\Big(\frac{1}{\Delta^2_{\min}} + \sum_{x\neq x^*} \frac{1}{\Delta^2(x)}\Big) = 2H_{\text{MAB}}. \end{aligned}$$

On the other hand, letting $\tilde{x}$ be the second best arm and $\Delta(x^*) = \Delta_{\text{min}}$, we have that

$$\begin{aligned} H_{\text{LB}} &= \min_{\lambda\in\mathcal{D}^k} \max_{x\neq x*} \frac{1/\lambda(x) + 1/\lambda(x^*)}{\Delta^2(x)} \\ &= \min_{\lambda\in\mathcal{D}^k} \max \left\{ \max_{x\neq x*} \frac{1/\lambda(x) + 1/\lambda(x^*)}{\Delta^2(x)}; \frac{1/\lambda(\tilde{x}) + 1/\lambda(x^*)}{\Delta^2(x^*)} \right\} \\ &\geq \min_{\lambda\in\mathcal{D}^k} \max \left\{ \max_{x\neq x*} \frac{1/\lambda(x)}{\Delta^2(x)}; \frac{1/\lambda(x^*)}{\Delta^2(x^*)} \right\} \\ &= \min_{\lambda\in\mathcal{D}^k} \max_{x\in\mathcal{X}} \frac{1/\lambda(x)}{\Delta^2(x)}. \end{aligned}$$

We set $\frac{1/\lambda(x)}{\Delta^2(x)}$ equal to a constant c and thus we get $\lambda(x) = \frac{1}{c\Delta^2(x)}$. Since $\frac{1}{c}\sum_{x\in\mathcal{X}} \frac{1}{\Delta^2(x)} = 1$, it follows that:

$$c = \sum_{x\in\mathcal{X}} \frac{1}{\Delta^2(x)} = \sum_{x\neq x^*} \frac{1}{\Delta^2(x)} + \frac{1}{\Delta^2_{\text{min}}} = H_{\text{MAB}}.$$

Thus, we get that $H_{\text{MAB}} \leq H_{\text{LB}} \leq 2H_{\text{MAB}}$. This shows that H_{LB} is a well defined notion of complexity for the linear best-arm identification problem and the corresponding sample complexity N^* is coherent with existing results in the MAB case. □

Proof of Lemma 3. The proof follows from the fact that if $\widehat{\mathcal{S}}(\mathbf{x}_n) \subseteq \mathcal{C}(x)$ and $\theta^* \in \widehat{\mathcal{S}}(\mathbf{x}_n)$ with high probability, then $\theta^* \in \mathcal{C}(x)$ which implies that $\Pi(\hat{\theta}_n) = x = x^*$. □

B.2 Proofs of Theorem 1 and Theorem 2

Proof of Theorem 1. The statement follows from Prop. 1 and the performance guarantees for the different implementations of the G-optimal design. By recalling the empirical stopping condition in Eq. 11 and the definition $\rho^G(\lambda) = \max_x x^\top \Lambda_\lambda^{-1} x$, we notice that from a simple triangle inequality applied to $||y||_{A^{-1}}$, a sufficient condition for stopping is that for any $x \in \mathcal{X}$

$$\frac{4c^2 \rho_n^{\tilde{G}} \log_n(K^2/\delta)}{n} \leq \widehat{\Delta}_n^2(x^*, x),$$

where $\rho_n^{\tilde{G}} = \rho^G(\lambda_{\mathbf{x}_n^{\tilde{G}}})$ and $\mathbf{x}_n^{\tilde{G}}$ is the allocation obtained from rounding the optimal design λ^G obtained from the continuous relaxation or the greedy incremental algorithm. From Prop. 1 we have that the following inequalities

$$\widehat{\Delta}_n(x^*, x) \geq \Delta(x^*, x) - c||x^* - x||_{A^{-1}_{\mathbf{x}_n^G}} \sqrt{\log_n(K^2/\delta)} \geq \Delta(x^*, x) - 2c\sqrt{\frac{\rho_n^{\tilde{G}} \log_n(K^2/\delta)}{n}},$$

hold with probability $1 - \delta$. Combining this with the previous condition and since the condition must hold for all $x \in \mathcal{X}$, we have that a sufficient condition to stop using the G-allocation is

$$\frac{16c^2 \rho_n^{\tilde{G}} \log_n(K^2/\delta)}{n} \leq \Delta_{\min},$$

which defines the level of accuracy that the G-allocation needs to achieve before stopping. Since $\rho_n^{\tilde{G}} \leq (1+\beta)d$ then the statement follows by inverting the previous inequality. □

Proof of Theorem 2. We follow the same steps as in the proof of Theorem 1. □

C Implementation of the Allocation Strategies

In this section we discuss about possible implementations of the allocation strategies illustrated in sections 4 and 5 and we discuss their approximation accuracy guarantees.

The efficient rounding procedure. We first report the general structure of the efficient rounding procedure defined in [15, Chapter 12] to implement a design λ into an allocation $\mathbf{x}_n$ for any fixed number of steps n. Let $p = \text{supp}(\lambda)$ the support of λ,[4] then we want to compute the number of pulls n_i (with $i = 1, \dots, p$) for all the arms in the support of λ. Basically, the fast implementation of the design is obtained in two phases, as follows:

- In the first phase, given the sample size n and the number of support points p, we calculate their corresponding frequencies $n_i = \lceil (n - \frac{1}{2}p)\lambda_i \rceil$, where $n_1, n_2, \dots, n_p$ are positive integers with $\sum_{i \leq p} n_i \geq n$.
- The second phase loops until the discrepancy $\left(\sum_{i \leq p} n_i\right) - n$ is 0, either:
  - increasing a frequency n_j which attains $n_j/\lambda_j = \min_{i \leq p}(n-1)/\lambda_i$ to n_{j+1}, or
  - decreasing some n_k with $(n_k - 1)/\lambda_k = \max_{i \leq p}(n_i - 1)/\lambda_i$ to $n - 1$.

An interesting feature of this procedure is that when moving from n to $n+1$ the corresponding allocations $\mathbf{x}_n$ and $\mathbf{x}_{n+1}$ only differ for one element i which is increased by 1, i.e., the discrete allocation is monotonic in n.

[4] For a fixed design $\lambda \in \mathbb{R}^K$, we say that its *support* is given by all arms in $\mathcal{X}$ whose corresponding features in λ are different than 0.

Implementation of the G-allocation. A first option is to optimize a continuous relaxation of the problem and compute the optimal design. Let $\rho^G(\lambda) = \max_x x^\top \Lambda_\lambda^{-1} x$, the optimal design is

$$\lambda^G = \arg\min_{\lambda \in \mathcal{D}_k} \max_{x \in \mathcal{X}} ||x||^2_{\Lambda_\lambda^{-1}} = \arg\min_{\lambda \in \mathcal{D}_k} \rho^G(\lambda). \tag{21}$$

This is a convex optimization problem and it can be solved using the projected gradient algorithm, interior point techniques, or multiplicative algorithms. To move from the design λ^G to a discrete allocation we use the efficient rounding technique presented above and we obtain that the resulting allocation $\mathbf{x}_t^{\tilde{G}}$ is guaranteed to be monotonic as the number of times an arm x is pulled is non-decreasing with t. Thus from $\mathbf{x}_t^{\tilde{G}}$ we obtain a simple incremental rule, where the arm x_t is the arm for which $\mathbf{x}_t^{\tilde{G}}$ recommends one pull more than in $\mathbf{x}_{t-1}^{\tilde{G}}$. An alternative is to directly implement an incremental version of Eq. 12 by selecting at each step t the greedy arm

$$x_t = \arg\min_{x \in \mathcal{X}} \max_{x' \in \mathcal{X}} x'^\top \big(A_{\mathbf{x}_{t-1}} + xx^\top\big)^{-1} x' = \arg\min_{x \in \mathcal{X}} \max_{x' \in \mathcal{X}} x'^\top \left[A^{-1}_{\mathbf{x}_{t-1}} - \frac{A^{-1}_{\mathbf{x}_{t-1}} x x^\top A^{-1}_{\mathbf{x}_{t-1}}}{1 + x^\top A^{-1}_{\mathbf{x}_{t-1}} x}\right] x', \tag{22}$$

where the second formulation follows from the matrix inversion lemma. This allocation is somehow simpler and more direct than using the continuous relaxation but it may come with a higher efficiency loss.

Before reporting the performance guarantees for the two implementations proposed above, we introduce an additional technical lemma which will be useful in the proofs on the performance guarantees. Although the lemma is presented for a specific definition of uncertainty ρ, any other notion including design matrices of the kind Λ_λ will satisfy the same guarantee.

Lemma 5. *Let $\rho(\lambda) = \max_{x \in \mathcal{X}} x^\top \Lambda_\lambda^{-1} x$ be a measure of uncertainty of interest for any design $\lambda \in \mathcal{D}^K$. We denote by $\lambda^* = \arg\min_{\lambda \in \mathcal{D}^K} \rho(\lambda)$ the optimal design and for any $n > d$ we introduce the optimal discrete allocation as*

$$\mathbf{x}_n^* = \arg\min_{\mathbf{x}_n \in \mathcal{X}^n} \max_{x \in \mathcal{X}} \frac{x^\top \Lambda^{-1}_{\lambda_{\mathbf{x}_n}} x}{n},$$

where $\lambda_{\mathbf{x}_n}$ is the (fractional) design corresponding to $\mathbf{x}_n$. Then we have

$$\rho(\lambda^*) \le \rho(\mathbf{x}_n^*) \le \Big(1 + \frac{p}{n}\Big)\rho(\lambda^*), \tag{23}$$

where $p = supp(\lambda^*)$ is the number of points in the support of λ^*. If d linearly independent arms are available in $\mathcal{X}$, then we can upper bound the size of the support of λ^* and obtain

$$\rho(\lambda^*) \le \rho(\mathbf{x}_n^*) \le \Big(1 + \frac{d(d+1)}{n}\Big)\rho(\lambda^*). \tag{24}$$

Proof. The first part of the statement follows by the definition of λ^* as the minimizer of ρ. Let $\tilde{\mathbf{x}}_n$ by an efficient rounding technique applied on λ^* such as the one described in Lemma 12.8 in [15]. Then $\tilde{\mathbf{x}}_n$ has the same support as λ^* and an efficiency loss bounded by p/n. As a result, we have

$$\rho(\mathbf{x}_n^*) \le \rho(\tilde{\mathbf{x}}_n) \le \Big(1 + \frac{p}{n}\Big)\rho(\lambda^*),$$

where the first inequality comes from the fact that $\mathbf{x}_n^*$ is the minimizer of ρ among allocations of length n. Then, from Caratheodory's theorem (see e.g., [15]) the number of support points used in λ^* is upper bounded by $p \le d(d+1)/2 + 1$ (under the assumption that there are d linearly independent arms in $\mathcal{X}$). The final result follows by a rough maximization of $d(d+1)/2n + 1/n \le d(d+1)/n$. □

Remark 1. Note that the same upper-bound for the number of support points holds for any design, due to the properties of the design matrices. In fact, any design matrix is symmetric by construction, which implies that it is completely described by $D = d(d+1)/2$ elements and can thus be seen as a point in $\mathbb{R}^D$. Moreover, a design matrix is a convex combination of a subset of points in $\mathbb{R}^D$ and thus it belongs to the convex hull of that subset of points. Caratheodory's theorem states that each point in the convex hull of any subset of points in $\mathbb{R}^D$ can be defined as a convex combination of at

most $D+1$ points. It directly follows that any design matrix can be expressed using $(d(d+1)/2)+1$ points.

It follows that the allocation $\mathbf{x}_t^{\tilde{G}}$ obtained applying the rounding procedure has the following performance guarantee.

Lemma 6. *For any $t \geq d$, the rounding procedure defined in [15, Chapter 12] returns an allocation $\mathbf{x}_t^{\tilde{G}}$, whose corresponding design $\lambda^{\tilde{G}} = \lambda_{\mathbf{x}_t^{\tilde{G}}}$ is such that*[5]

$$\rho^G(\lambda^{\tilde{G}}) \leq \Big(1 + \frac{d + d^2 + 2}{2t}\Big) d.$$

Proof of Lemma 6. We follow the same steps as in the proof of Lemma 5 to obtain the term $\beta = \frac{d+d^2+2}{2t}$. Then, noting that the performance of the optimal strategy $\rho^G(\lambda^{*G}) = d$ (from Prop. 3), the results follows. □

Implementation of the $\mathcal{XY}$-allocation. Notice that the complexity of the $\mathcal{XY}$-allocation trivially follows from the complexity of the G-allocation and it is NP-hard. As a result, we need to propose approximate solutions to compute an allocation $\mathbf{x}_n^{\widetilde{\mathcal{XY}}}$ as for the G-allocation. Let $\rho^{\mathcal{XY}}(\lambda) = \max_{y \in \mathcal{Y}} y^\top \Lambda_\lambda^{-1} y$, then the first option is the compute the optimal solution to the continuous relaxed problem

$$\lambda^{\mathcal{XY}} = \arg\min_{\lambda \in \mathcal{D}_k} \max_{y \in \mathcal{Y}} ||y||^2_{\Lambda_\lambda^{-1}} = \arg\min_{\lambda \in \mathcal{D}_k} \rho^{\mathcal{XY}}(\lambda). \tag{25}$$

And then compute the corresponding discrete allocation $\mathbf{x}_n^{\widetilde{\mathcal{XY}}}$ using the efficient rounding procedure. Alternatively, we can use an incremental greedy algorithm which at each step t returns the arm

$$x_t = \arg\min_{x \in \mathcal{X}} \max_{y \in \mathcal{Y}} y^\top \big(A_{\mathbf{x}_{t-1}} + xx^\top\big)^{-1} y. \tag{26}$$

Lemma 7. *For any $t \geq d$, the rounding procedure defined in [15, Chapter 12] returns an allocation $\mathbf{x}_t^{\widetilde{\mathcal{XY}}}$, whose corresponding design $\lambda^{\widetilde{\mathcal{XY}}} = \lambda_{\mathbf{x}_t^{\widetilde{\mathcal{XY}}}}$ is such that*

$$\rho^{\mathcal{XY}}(\lambda^{\widetilde{\mathcal{XY}}}) \leq 2\Big(1 + \frac{d + d^2 + 2}{2t}\Big) d.$$

Proof of Lemma 7. The proof follows from the fact that for any pair (x, x')

$$||x - x'||_{A^{-1}_{\mathbf{x}_n}} \leq 2 \max_{x'' \in \mathcal{X}} ||x''||_{A^{-1}_{\mathbf{x}_n}}.$$

Then the proof proceeds as in Lemma 6. □

Implementation of $\mathcal{XY}$-adaptive allocation. The allocation rule in Eq. 17 basically coincides with the $\mathcal{XY}$-allocation and its properties extend smoothly.

D Proof of Theorem 3

Before proceeding to the proof, we first report the proofs of two adittional lemmas.

Proof of Lemma 4. Let $y = x' - x$. Using the definition of $\widehat{\mathcal{S}}(\mathbf{x}_n)$ in Eq. 10, and the fact that $\theta^* \in \widehat{\mathcal{S}}(\mathbf{x}_n)$ with high probability, we have

$$(x' - x)^\top (\hat{\theta}_n - \theta^*) \leq c||x' - x||_{A^{-1}_{\mathbf{x}}} \sqrt{\log_n(K^2/\delta)}.$$

Since the condition in Eq. 16 is true, it follows that

$$\begin{aligned}(x' - x)^\top (\hat{\theta}_n - \theta^*) \leq c||x' - x||_{A^{-1}_{\mathbf{x}}} \sqrt{\log_n(K^2/\delta)} \leq \widehat{\Delta}_n(x', x) \Leftrightarrow \\ -(x' - x)^\top \theta^* \leq 0 \Leftrightarrow x^\top \theta^* \leq x'^\top \theta^*\end{aligned}$$

thus x is dominated by x' and x cannot be the optimal arm. □

[5] We recall that from any allocation $\mathbf{x}_n$ the corresponding design $\lambda_{\mathbf{x}}$ is such that $\lambda_{\mathbf{x}_n}(x) = T_n(x)/n$.

Lemma 8. *For any phase j, the length is such that $n_j \leq \max\{M^*, \frac{16}{\alpha}N^*\}$ with probability $1-\delta$.*

Proof of Lemma 8. We first summarize the different quantities measuring the performance of an allocation strategy in different settings. For any design $\lambda \in \mathcal{D}^K$, we define

$$\rho^*(\lambda) = \max_{y \in \mathcal{Y}^*} \frac{||y||^2_{\Lambda_\lambda^{-1}}}{\Delta^2(y)}; \quad \rho^{\mathcal{XY}}(\lambda) = \max_{y \in \mathcal{Y}} ||y||^2_{\Lambda_\lambda^{-1}}; \quad \rho^j(\lambda) = \max_{y \in \widehat{\mathcal{Y}}_j} ||y||^2_{\Lambda_\lambda^{-1}}. \tag{27}$$

For any n, we also introduce the value of each of the previous quantities when the corresponding optimal (discrete) allocation is used

$$\rho_n^* = \rho^*(\lambda_{\mathbf{x}_n^*}); \quad \rho_n^{\mathcal{XY}} = \rho^{\mathcal{XY}}(\lambda_{\mathbf{x}_n^{\mathcal{XY}}}); \quad \rho_n^j = \rho^j(\lambda_{\mathbf{x}_n^j}). \tag{28}$$

Finally, we introduce the optimal designs

$$\lambda^* = \arg\min_{\lambda \in \mathcal{D}^K} \rho^*(\lambda); \;\; \lambda^{\mathcal{XY}} = \arg\min_{\lambda \in \mathcal{D}^K} \rho^{\mathcal{XY}}(\lambda); \;\; \lambda^j = \arg\min_{\lambda \in \mathcal{D}^K} \rho^j(\lambda). \tag{29}$$

Let ϵ^* be the smallest ϵ such that there exists a pair (x, x'), with $x \neq x^*$ and $x' \neq x^*$, such that the confidence set $\mathcal{S} = \{\theta : \forall y \in \mathcal{Y}, |y^\top(\theta - \theta^*)| \leq \epsilon\}$ overlaps with the hyperplane $\mathcal{C}(x) \cap \mathcal{C}(x')$. Since M^* is defined as the smallest number of steps needed by the $\mathcal{XY}$ strategy to avoid any overlap between $\mathcal{S}^*$ and the hyperplanes $\mathcal{C}(x) \cap \mathcal{C}(x')$, then we have that after M^* steps

$$c\sqrt{\frac{\rho_{M^*}^{\mathcal{XY}} \log_n(K^2/\delta)}{M^*}} < \epsilon^*. \tag{30}$$

We consider two cases to study the length of a phase j.

Case 1: $\sqrt{\frac{\rho_{n_j}^j}{n_j}} \geq \frac{\epsilon^*}{c\sqrt{\log_n(K^2/\delta)}}$. From Eq. 30 it immediately follows that

$$\frac{\rho_{n_j}^j}{n_j} \geq \frac{\rho_{M^*}^{\mathcal{XY}}}{M^*}. \tag{31}$$

From definitions in Eqs. 27 and 28, since $\widehat{\mathcal{Y}}_j \subseteq \mathcal{Y}$ we have for any n, $\rho_n^j \leq \rho_n^{\mathcal{XY}}$. As a result, if $n_j \geq M^*$, since ρ_n^j/n is a non-increasing function, then we would have the sequence of inequalities

$$\frac{\rho_{n_j}^j}{n_j} \leq \frac{\rho_{M^*}^j}{M^*} \leq \frac{\rho_{M^*}^{\mathcal{XY}}}{M^*},$$

which contradicts Eq. 31. Thus $n_j \leq M^*$.

Case 2: $\sqrt{\frac{\rho_{n_j}^j}{n_j}} \leq \frac{\epsilon^*}{c\sqrt{\log_n(K^2/\delta)}}$. We first relate the performance at phase j with the performance of the oracle. For any n

$$\rho_n^j = \rho^j(\lambda_{\mathbf{x}_n^j}) \leq \rho^j(\lambda_{\mathbf{x}_n^*}) = \max_{y \in \widehat{\mathcal{Y}}_j} y^\top \Lambda^{-1}_{\lambda_{\mathbf{x}_n^*}} y = \max_{y \in \widehat{\mathcal{Y}}_j} \frac{y^\top \Lambda^{-1}_{\lambda_{\mathbf{x}_n^*}} y}{\Delta^2(y)} \Delta(y) \leq \max_{y \in \widehat{\mathcal{Y}}_j} \frac{y^\top \Lambda^{-1}_{\lambda_{\mathbf{x}_n^*}} y}{\Delta^2(y)} \max_{y \in \widehat{\mathcal{Y}}_j} \Delta^2(y).$$

If now we consider $n = n_j$, then the definition case 2 implies that the estimation error $\sqrt{\rho_{n_j}^j/n_j}$ is small enough so that all the directions in $\mathcal{Y} - \mathcal{Y}^*$ have already been discarded from $\widehat{\mathcal{Y}}_j$ and $\widehat{\mathcal{Y}}_j \subseteq \mathcal{Y}^*$. Thus

$$\rho_{n_j}^j \leq \max_{y \in \mathcal{Y}^*} \frac{y^\top \Lambda^{-1}_{\lambda_{\mathbf{x}_{n_j}^*}} y}{\Delta^2(y)} \max_{y \in \widehat{\mathcal{Y}}_j} \Delta^2(y) = \rho_{n_j}^* \max_{y \in \widehat{\mathcal{Y}}_j} \Delta^2(y). \tag{32}$$

This relationship does not provide a bound on n_j yet. We first need to recall from Prop. 1 that for any $y \in \mathcal{Y}$ (and notably for the directions in $\widehat{\mathcal{Y}}_j$) we have

$$|y^\top(\hat{\theta}_{j-1} - \theta^*)| \leq c\sqrt{y^\top A_{j-1}^{-1} y \log_n(K^2/\delta)},$$

where $A_{j-1} = A_{\mathbf{x}_{n_{j-1}}^{j-1}}$ is the matrix constructed from the pulls within phase $j-1$. Since $\mathbf{x}_n^{j-1}$ is obtained from a $\mathcal{XY}$-allocation applied on directions in $\widehat{\mathcal{Y}}_{j-1}$, we obtain that for any $y \in \widehat{\mathcal{Y}}_j$

$$|y^\top(\hat{\theta}_{j-1} - \theta^*)| \le c\sqrt{\log_n(K^2/\delta)} \max_{y \in \widehat{\mathcal{Y}}_{j-1}} \sqrt{y^\top A_{j-1}^{-1} y} = c\sqrt{\frac{\log_n(K^2/\delta)\rho_{n_{j-1}}^{j-1}}{n_{j-1}}},$$

Reordering the terms in the previous expression we have that for any $y \in \widehat{\mathcal{Y}}_j$

$$\Delta(y) \le \widehat{\Delta}_{j-1}(y) + c\sqrt{\frac{\log_n(K^2/\delta)\rho_{n_{j-1}}^{j-1}}{n_{j-1}}}.$$

Since the direction y is included in $\widehat{\mathcal{Y}}_j$ then the discard condition in Eq. 16 failed for y, implying that $\widehat{\Delta}_{j-1}(y) \le c\sqrt{\frac{\log_n(K^2/\delta)\rho_{n_{j-1}}^{j-1}}{n_{j-1}}}$. Thus we finally obtain

$$\max_{y \in \widehat{\mathcal{Y}}_j} \Delta(y) \le 2c\sqrt{\frac{\log_n(K^2/\delta)\rho_{n_{j-1}}^{j-1}}{n_{j-1}}}.$$

Combining this with Eq. 32 we have

$$\rho_{n_j}^j \le \rho_{n_j}^* 4c^2 \frac{\log_n(K^2/\delta)\rho_{n_{j-1}}^{j-1}}{n_{j-1}}.$$

Using the stopping condition of phase j and the relationship between the performance ρ^j, we obtain that at time $\bar{n} = n_j - 1$

$$\frac{\rho_{\bar{n}}^j}{\bar{n}} \ge \alpha \frac{\rho_{n_{j-1}}^{j-1}}{n_{j-1}} \ge \frac{\alpha}{4c^2 \log_n(K^2/\delta)} \frac{\rho_{n_j}^j}{\rho_{n_j}^*}$$

We can further refine the previous inequality as

$$\frac{\rho_{\bar{n}}^j}{\bar{n}} \ge \frac{\alpha\rho_{N^*}^*}{4N^*} \frac{N^*}{c^2 \log_n(K^2/\delta)\rho_{N^*}^*} \frac{\rho_{n_j}^j}{\rho_{n_j}^*} \ge \frac{\alpha\rho_{N^*}^*}{4N^*} \frac{\rho_{n_j}^j}{\rho_{n_j}^*},$$

where we use the definition of N^* in Eq. 7, which implies $c\sqrt{\log_n(K^2/\delta)\rho_{N^*}^*/N^*} \le 1$. Reordering the terms and using $\bar{n} = n_j - 1$, we obtain

$$n_j \le 1 + \frac{4N^*}{\alpha} \frac{\rho_{n_j-1}^j}{\rho_{n_j}^j} \frac{\rho_{n_j}^*}{\rho_{N^*}^*}.$$

From Lemma 5 and the optimal designs defined in Eq. 29 we have

$$n_j \le 1 + \frac{4N^*}{\alpha} \frac{(1 + d(d+1)/(n_j-1))\rho^j(\lambda^j)}{\rho^j(\lambda^j)} \frac{(1 + d(d+1)/(n_j-1))\rho^*(\lambda^*)}{\rho^*(\lambda^*)}.$$

Using the fact that the algorithm forces $n_j \ge d(d+1)+1$, the statement follows. □

Proof of Theorem 3. Let J be the index of any phase for which $|\widehat{\mathcal{X}}_J| > 1$. Then there exist at least one arm $x \in \mathcal{X}$ (beside x^*) for which the discarding condition in Lemma 4 is not triggered, which corresponds to the fact that for all arms $x' \in \mathcal{X}$

$$c||x - x'||_{A_{\mathbf{x}_{n_J}^J}^{-1}} \sqrt{\log_n(K^2/\delta)} \ge \widehat{\Delta}_J(x, x').$$

By developing the right hand side, we have

$$\widehat{\Delta}_J(x, x') \ge \Delta(x, x') - c||x - x'||_{A_{\mathbf{x}_{n_J}^J}^{-1}} \sqrt{\log_n(K^2/\delta)} \ge \Delta_{\min} - c\sqrt{\frac{\rho_{n_J}^J \log_n(K^2/\delta)}{n_J}}$$

which leads to the condition

$$2c\sqrt{\frac{\rho_{n_J}^J \log_n(K^2/\delta)}{n_J}} \geq \Delta_{\min}. \tag{33}$$

Using the phase stopping condition and the initial value of ρ^0 we have

$$\frac{\rho_{n_J}^J}{n_J} \leq \alpha \frac{\rho_{n_{J-1}}^{J-1}}{n_{J-1}} \leq \alpha^J \frac{\rho^0}{n_0} = \alpha^J.$$

By joining this inequality with Eq. 33 we obtain

$$\alpha^J \geq \frac{\Delta_{\min}^2}{4c^2 \log_n(K^2/\delta)},$$

and it follows that $J \leq \log(4c^2 \log_n(K^2/\delta)/\Delta_{\min}^2)/\log(1/\alpha)$ which together with Lemma 8 leads to the final statement. □

E Additional Empirical Results

For the setting described in Sec. 6, in order to point out the different repartitions of the sampling budget over arms, in Fig. 5 we present the number of samples allocated per arm, for the case when the input space $\mathcal{X} \subseteq \mathbb{R}^5$. We remind that the arms denoted $x_1, \ldots, x_5$ form the canonical basis and arm $x_6 = [\cos(\omega) \;\; \sin(\omega) \;\; 0 \;\; 0 \;\; 0]$.

Samples/arm	*$\mathcal{XY}$-oracle*	*$\mathcal{XY}$-adaptive*	*$\mathcal{XY}$*	*G*	*Fully-adaptive*
x_1	207	263	29523	28014	740
x_2	41440	52713	29524	28015	149220
x_3	2	3	29524	28015	1
x_4	2	5	29524	28015	1
x_5	1	2	29524	28015	1
x_6	0	2	1	1	1
Budget	41652	52988	147620	140075	149964

Figure 5: The budget needed by the allocation strategies to identify the best arm when $\mathcal{X} \subseteq \mathbb{R}^5$ and their sample allocation over arms. $\mathcal{XY}$ and G allocate samples uniformly over the canonical arms while $\mathcal{XY}$-oracle and $\mathcal{XY}$-adaptive use most of the samples for arm x_2 (corresponding to the most informative direction).

We can notice that even though the Fully-adaptive algorithm identifies the most informative direction and focuses the sampling on arm x_2, its sample complexity still has a growth linear in the dimension, due to the extra $\sqrt{d}$ term in his bound. Consequently, the advantage over the static strategies is canceled. On the other hand, $\mathcal{XY}$-adaptive "learns" the gaps from the observations and allocates the samples very similarly to $\mathcal{XY}$-oracle, without suffering a large loss in terms of the sampling budget. However, $\mathcal{XY}$-adaptive's sample complexity has to account for the the re-initializations made at the beginning of a new phase.

Finally, we notice that in this problem that static allocations, $\mathcal{XY}$ and G, perform a uniform allocation over the canonical arms. Another interesting remark is that the number of pulls to one canonical arm is smaller than the samples that $\mathcal{XY}$-oracle allocated to x_2. This is explained by the "mutual information" coming from the multiple observations on all directions, which helps in reducing the overall uncertainty of the confidence set.